MULTIMEDIA UNIVERSITY

STUDENT ID NO

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 1, 2016/2017

TMM 3341– DATA MINING AND MACHINE LEARNING SYSTEMS

(All sections / Groups)

22 OCTOBER 2016
9.00am – 11.00am
(2 Hours)

INSTRUCTIONS TO STUDENT

1. This Question paper consists of 4 pages including cover page.
2. Answer all **FIVE** questions. All questions carry equal marks and the distribution of the marks for each question is given.
3. Please write all your answers in the Answer Booklet provided.

QUESTION 1

a. What roles does exploratory data analysis play in a data mining project? List **THREE (3)** roles. (3 marks)

b. Describe the **THREE (3)** main differences between Classification and clustering tasks? (3 marks)

c. What data mining function does a department store need to assist with its target marketing mail campaign? Explain your answer. (4 marks)

QUESTION 2

a. Explain **THREE (3)** characteristic of data quality. (3 marks)

b. Generally, classification is a two-step process. Explain these **TWO (2)** steps. What are the roles of the training-set and test-set in this process? (4 marks)

c. MGS collage has about 26,000 students. Every year, 1.4% of all students attempt to change faculty, and there are about 2 students who actual change faculty. Students who have attempted to change faculty are 543 times more likely to change faculty successfully in the next year.

   i. What is the percentage of all successful change faculty attempts? (1 mark)

   ii. Now, suppose we have a database with detailed information about each MGS student, with two labels for each person: student actually change faculty and student attempted to change faculty. Explain why it would or would not be useful to train a classifier to distinguish the actual change of faculty from all other students. (2 marks)

QUESTION 3

a. Suppose a group of 12 students with the test scores listed as follows: 74, 31, 64, 85, 19, 88, 72, 79, 71, 48, 99, and 95. Partition them into four bins by

   i. equal-frequency (equi-depth) method (1 mark)

   ii. equal-width method (2 marks)

b. In real world data, tuples with missing values for some attributes are a common occurrence. Describe **TWO (2)** methods for handling this problem. (4 marks)

c. Explain **TWO (2)** needs of machine learning in data mining? (3 marks)

Continued .....

QUESTION 4

a. Consider the training dataset in Table 1 which consist of 12 data instances:

Table 1

Age	Spectacle Prescription	Astigmatism	Tear Production Rate	Contact Lenses
pre-presbyopic	myope	no	normal	soft
young	myope	no	normal	soft
presbyopic	hypermetrope	no	normal	soft
presbyopic	myope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
young	myope	yes	normal	hard
presbyopic	myope	no	normal	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	myope	yes	reduced	none
young	hypermetrope	no	reduced	none
pre-presbyopic	myope	no	reduced	none

i. Let Contact Lenses be the classification target. Construct the ZeroR model for the training dataset. Explain your work. (2 marks)

ii. What Contact Lenses class does ZeroR predict for the test data instance? (1 mark)

b. A naive Bayes classifier was trained on a dataset consisting of spam and non-spam emails. The result of the trained classifier for 10 instances is shown in Table 2. Give the confusion matrix and misclassification rate when we use a prediction threshold of 0.990 (i.e., we classify an instance as spam if the predicted probability is greater than 0.990). (5 marks)

Table 2

Predicted Probability	Actual Label
0.001	not spam
0.100	not spam
0.500	not spam
0.600	not spam
0.980	not spam
0.400	spam
0.880	spam
0.900	spam
0.995	spam
0.999	spam

c. Assume the itemset {A, B, D} is frequent; what rules will be generated from this itemset? (2 marks)

Continued .....

QUESTION 5

a. There are **FOUR (4)** errors in the following .arff file. Correct the errors. Write only the line numbers and the corrected lines. You do not have to rewrite the entire file. (4 marks)

Table 2

1	relation movie-preferences
2	@attribute genre {comedy, drama, action}
3	@attribute critics-reviews {good, bad}
4	@attribute windy {TRUE, FALSE}
5	@attribute play yes, no
6	@movie
7	drama, good, FALSE, no
8	comedy, good, TRUE, no
9	comedy, neutral, FALSE, yes
10	action, bad, FALSE, yes

b. Suppose a text retrieval system has retrieved a number of documents for Azman based on his query. Describe **ONE (1)** method of how Ali can assess the quality of text retrieval system? (2 marks)

c. A company asks you for help. They want to select the best clustering algorithm for their data, among a set of many algorithms. They can provide you with a dataset that consists of data (features) about their customers. The dataset has been labeled by a company expert, who labeled each customer record as "HIGH" or "LOW" based on the customer's spending level. How would you use this data to perform effective clustering (i.e., what parts of the data would you use for training the algorithm and what parts would you use for testing it)? (4 marks)

End of Page